FermiGrid Scalability and Reliability Improvements

K. Chadwick, F. Lowe, N. Sharma, S. Timm, D. R. Yocum
Grid And Cloud Computing Department
Fermilab
ISGC2011

Work supported by the U.S. Department of Energy under contract No. DE-AC02-07CH11359

Fermilab

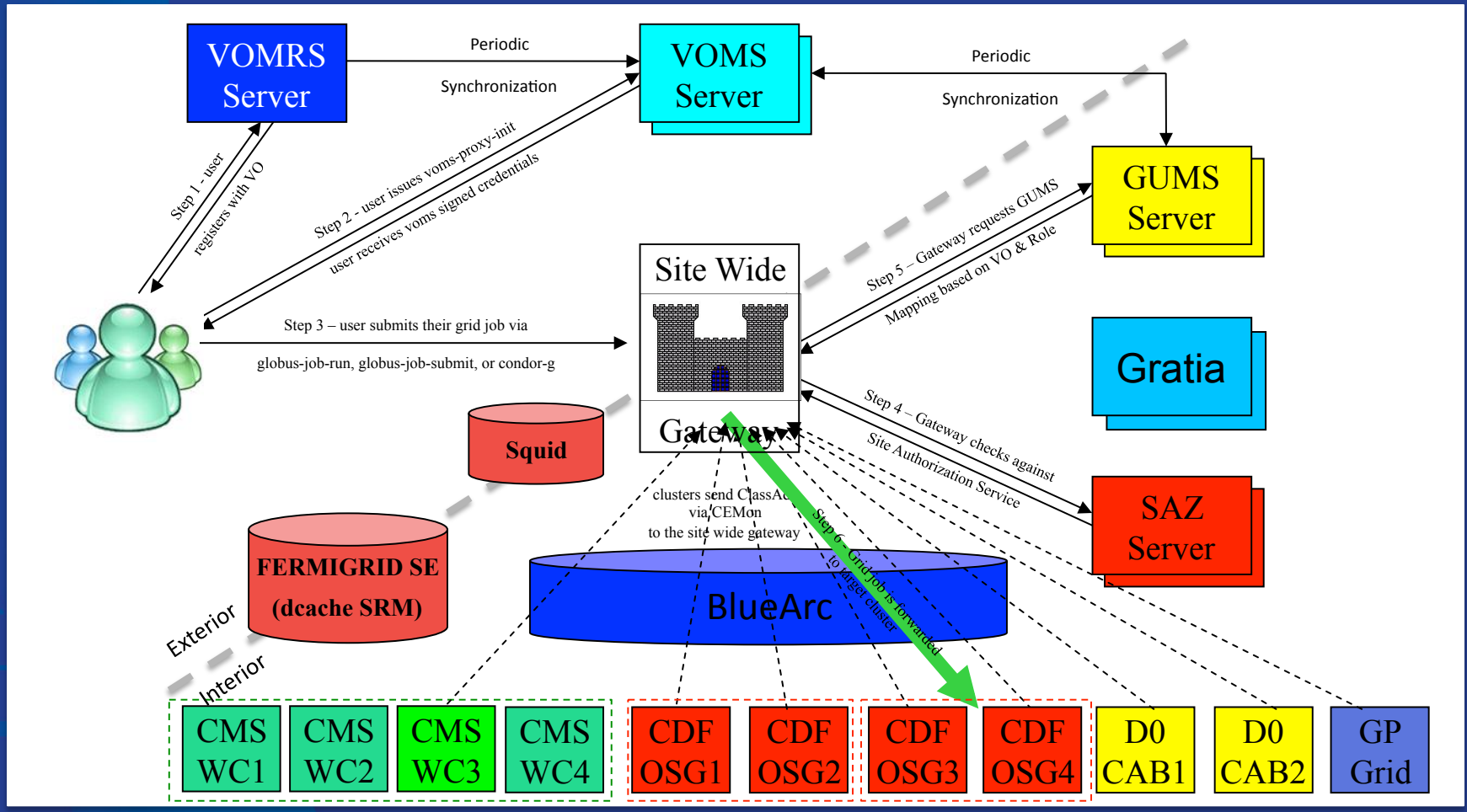
What is FermiGrid?

- A Meta-facility that provides grid infrastructure for scientific computing at Fermilab,
- Provides highly-available centralized authorization and authentication services,
- Provides site gateway for Globus job submission,
- Coordinates interoperability among stakeholders,
- Provides grid-enabled mass storage services,
- 2.1x10^8 CPU-hours recorded since Gratia accounting started, most in OSG,
- FermiGrid now has ~22500 batch slots, 10 compute elements.

Key Policies

- Early management directions:
  - Special security enclave for nodes where grid jobs run,
  - There WILL be a single site gateway for OSG jobs,
  - All major users and clusters MUST interoperate,
  - There WILL be a unified Grid Cert -> Unix uid mapping,
  - There WILL be a central banning server,
  - Pilot jobs MUST use gLexec to authenticate 3rd-party jobs,
  - We WILL have our own Certificate Authority (the Fermilab Kerberos Certificate Authority),
  - We WILL keep more than one batch system in production.

FermiGrid Architecture



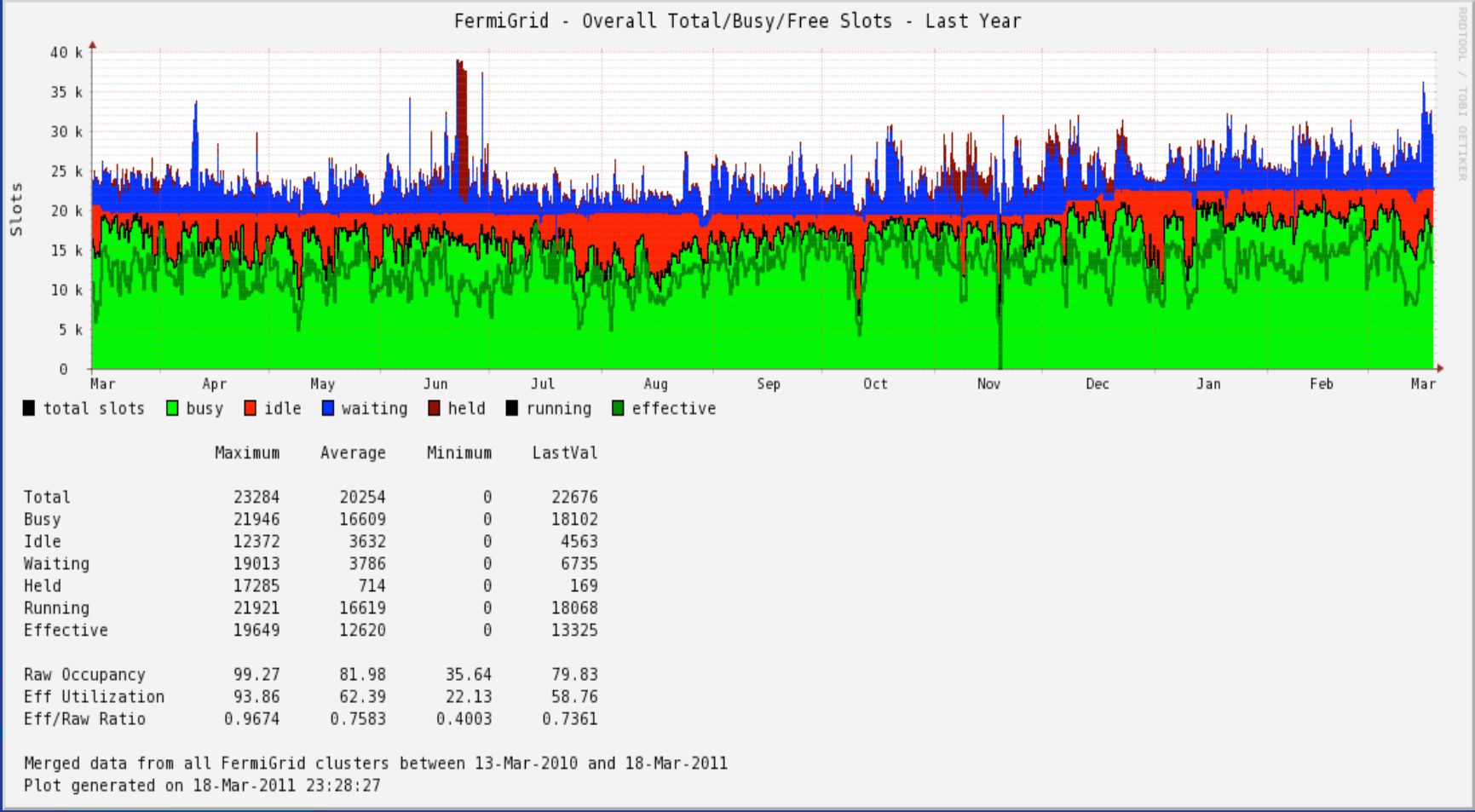
Condor Classads

- Start with Glue Schema 1.3 LDIF information for each Computing Element and associated Storage Elements,
- Clusters divided up into subclusters based on hardware type,
- gLite CEMon plugin makes one classad for each unique combination of {VO, subcluster, Storage Area},
- Example: 60 VO's x 4 subclusters x 2 Storage Areas = 480 classads in just 1 cluster,
- FermiGrid presently has 7,174 classads,
- OSG presently has 21,981 classads,
- Classads transmitted by CEMon to Resource Selection System,
- Raw LDIF transmitted by CEMon to OSG BDII.

Recent use case - OS Version

- Needed to upgrade from SLF4 to SLF5, one subcluster at a time,
- Each subcluster advertises:
  - Current OS Version,
  - Extra RSL fields to make sure you run on those nodes when your job matches to that subcluster.
- Users submit job requesting OS (4.7 or 5.3),
- Site gateway finds a matching subcluster and automatically appends right RSL, and forwards job.
- Can also select on memory, disk, batch system.

Overall Occupancy & Utilization



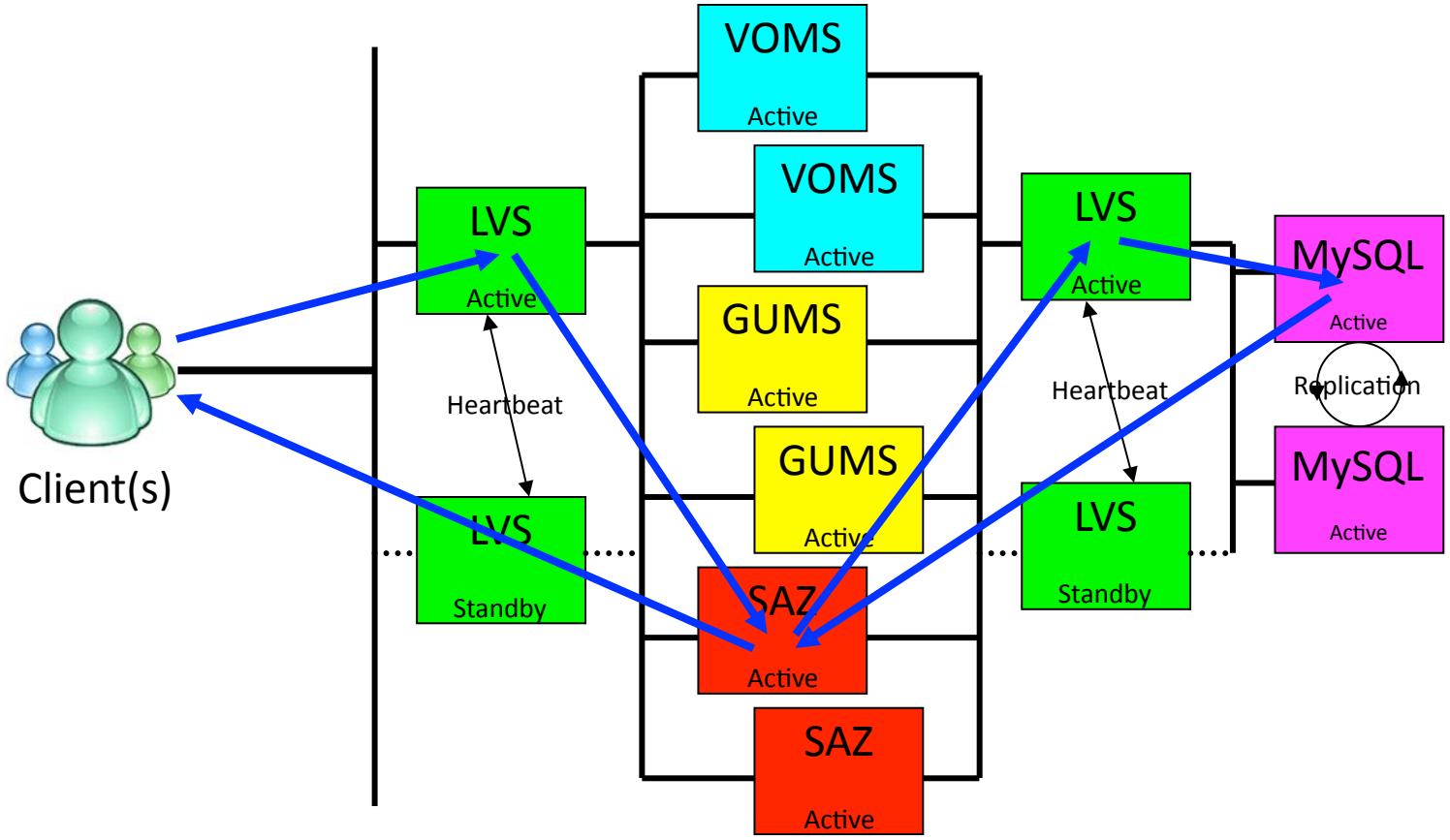
	Maximum	Average	Minimum	LastVal
Total	23284	20254	0	22676
Busy	21946	16609	0	18102
Idle	12372	3632	0	4563
Waiting	19013	3786	0	6735
Held	17285	714	0	169
Running	21921	16619	0	18068
Effective	19649	12620	0	13325
Raw Occupancy	99.27	81.98	35.64	79.83
Eff Utilization	93.86	62.39	22.13	58.76
Eff/Raw Ratio	0.9674	0.7583	0.4003	0.7361

Merged data from all FermiGrid clusters between 13-Mar-2010 and 18-Mar-2011
Plot generated on 18-Mar-2011 23:28:27

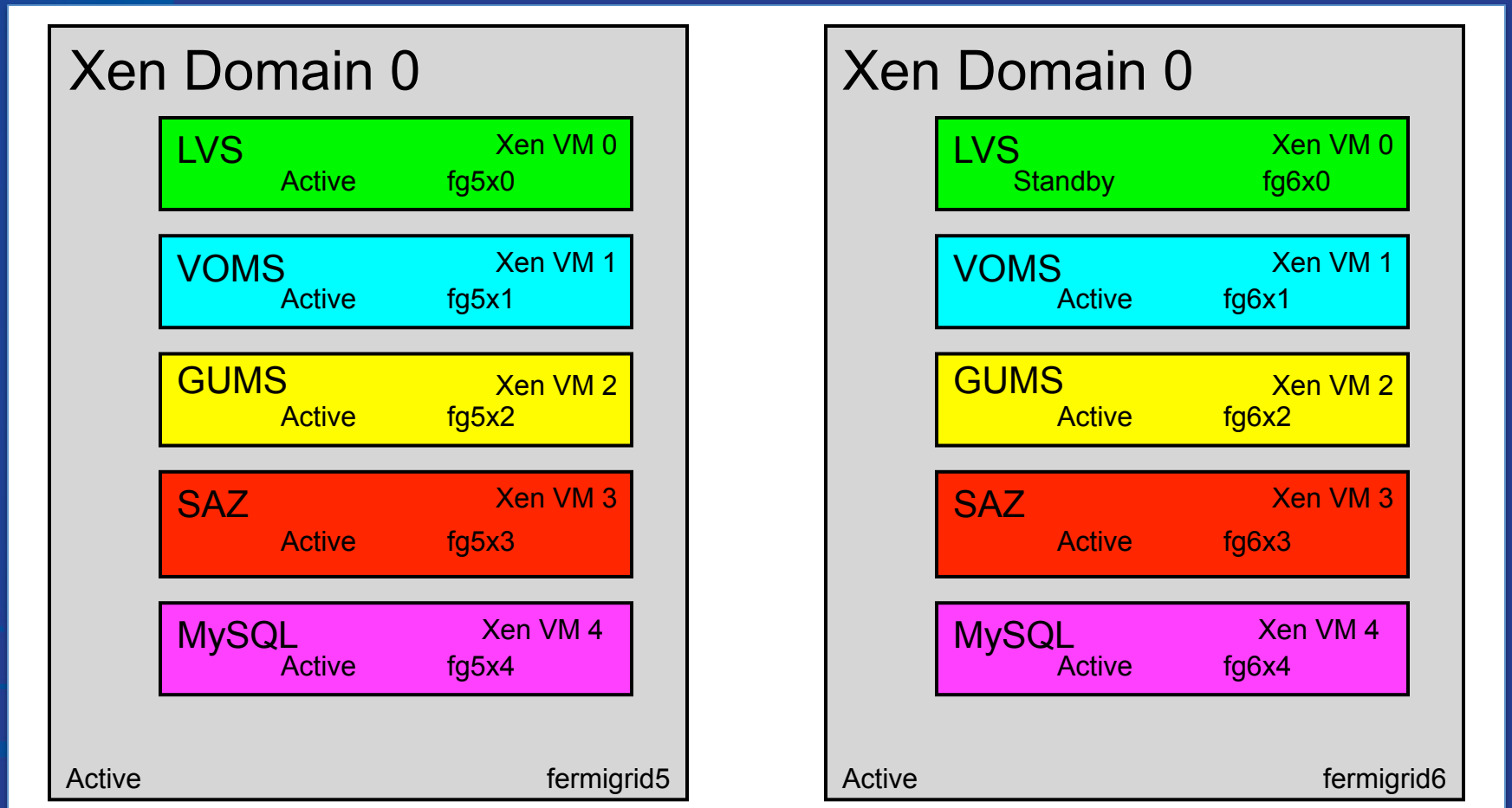
Batch Systems, Occupancy & Utilization

Cluster	Cluster Batch System	Current Cluster Size (Slots)	Average Cluster Occupancy (%)	Average Cluster Utilization (%)
CDF	Condor	5,600	89.2	61.7
CMS	Condor	7,485	86.3	73.9
D0	PBS	6,916	74.8	49.6
GP	Condor	3,284	69.0	63.0
Overall	----	23,285	82.0	62.4

FermiGrid HA Services - 1



FermiGrid-HA Services - 2



Measured FermiGrid Service Availability for the Past Year*

Service	Availability	Downtime
VOMS-HA	100%	0m
GUMS-HA	100%	0m
SAZ-HA (gatekeeper)	100%	0m
Squid-HA	100%	0m
MyProxy-HA	99.943%	299.0m
ReSS-HA	99.959%	215.36m
Gratia-HP	99.955%	233.30m
Database-HA	99.963%	192.62m

* = Excluding building or network failures

SAZ - Central Banning Service

- Site AuthoriZation service, developed at Fermilab,
- Allows us to ban any user, VO, CA, group, or role,
- Don't have to wait for CA to revoke the cert or VO to remove from membership,
- Available as Pacman package,
- Details at http://saz.fnal.gov
- With glideins, every worker node can be a client simultaneously (have seen 5000+ active clients),
- Significant work has been done to make it more resilient and scalable,
- Code also contains support for new XACML-based authorization protocol.

Why a new SAZ Server?

- Previous SAZ server (V2_0_1b) has shown itself extremely vulnerable to user generated authorization "tsunamis":
  - Very short duration jobs
  - User issues condor_rm on a large (>1000) glidein.
- This is fixed in the new SAZ Server (V2_7_2) using tomcat and pools of execution and hibernate threads.
- Various other bugs were found and fixed in the current SAZ server and sazclient.
- Added support for the XACML protocol (used by Globus).
  - NOT transitioning to using XACML (yet).

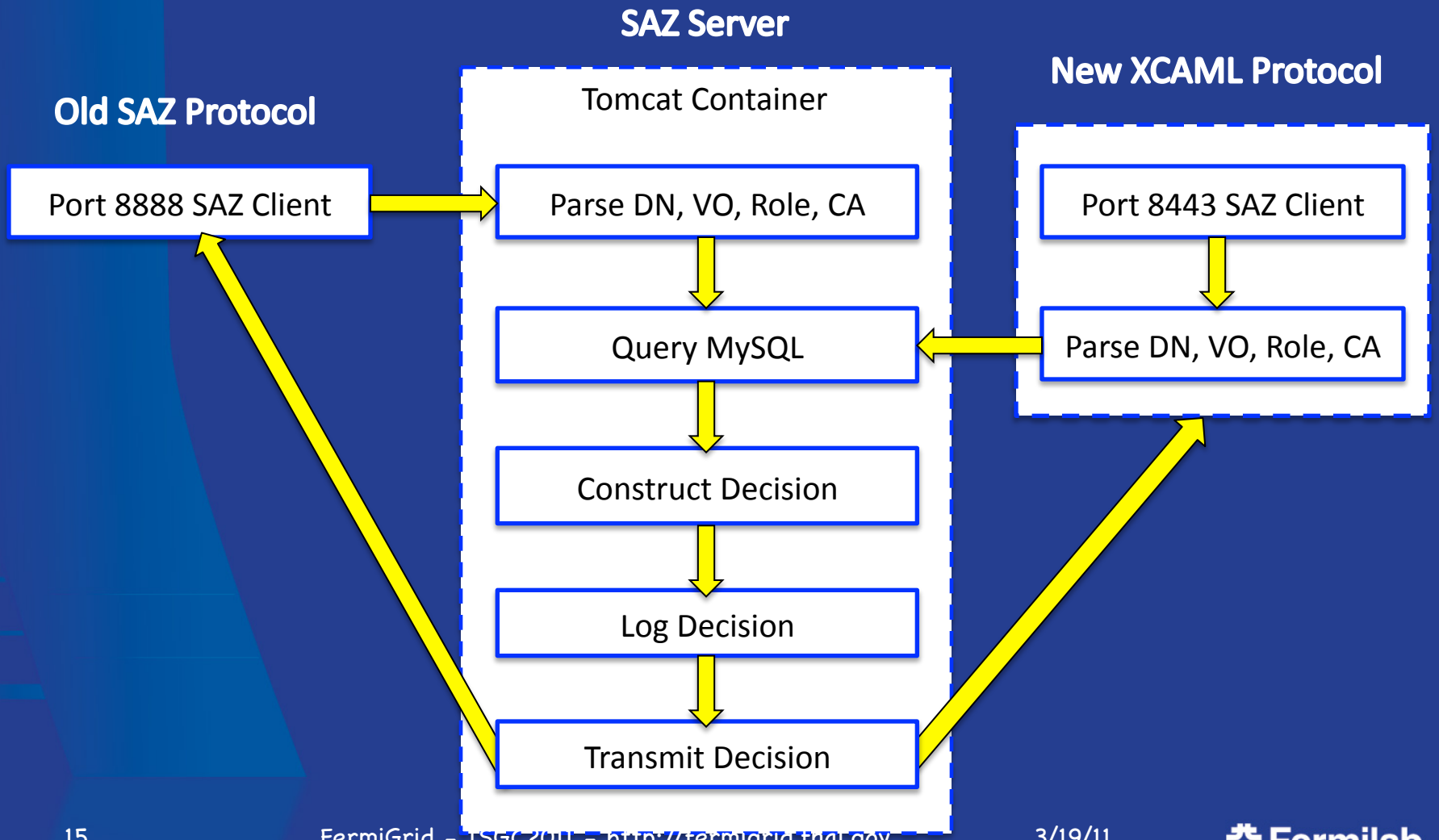
Old ("current") SAZ Protocol - Port 8888

- Client sends the "entire" proxy to the SAZ server via port 8888.
- Server parses out DN, VO, Role, CA.
  - In SAZ V2.0.0b, the parsing logic does not work well, and frequently the SAZ server has to invoke a shell script voms-proxy-info to parse the proxy.
  - In the new SAZ V2_7_2, the parsing logic has been completely rewritten, and it no longer has to invoke the shell script voms-proxy-info to parse the proxy.
- Server performs MySQL queries.
- Server constructs the answer and sends it to the client.

New SAZ (XACML) Protocol – Port 8443

- Client parses out DN, VO, Role, CA and sends the information via XACML to the SAZ server via port 8443.
- Server performs MySQL queries.
- Server constructs the answer and sends it to the client.
- The new SAZ server supports both 8888 and 8443 protocols simultaneously.

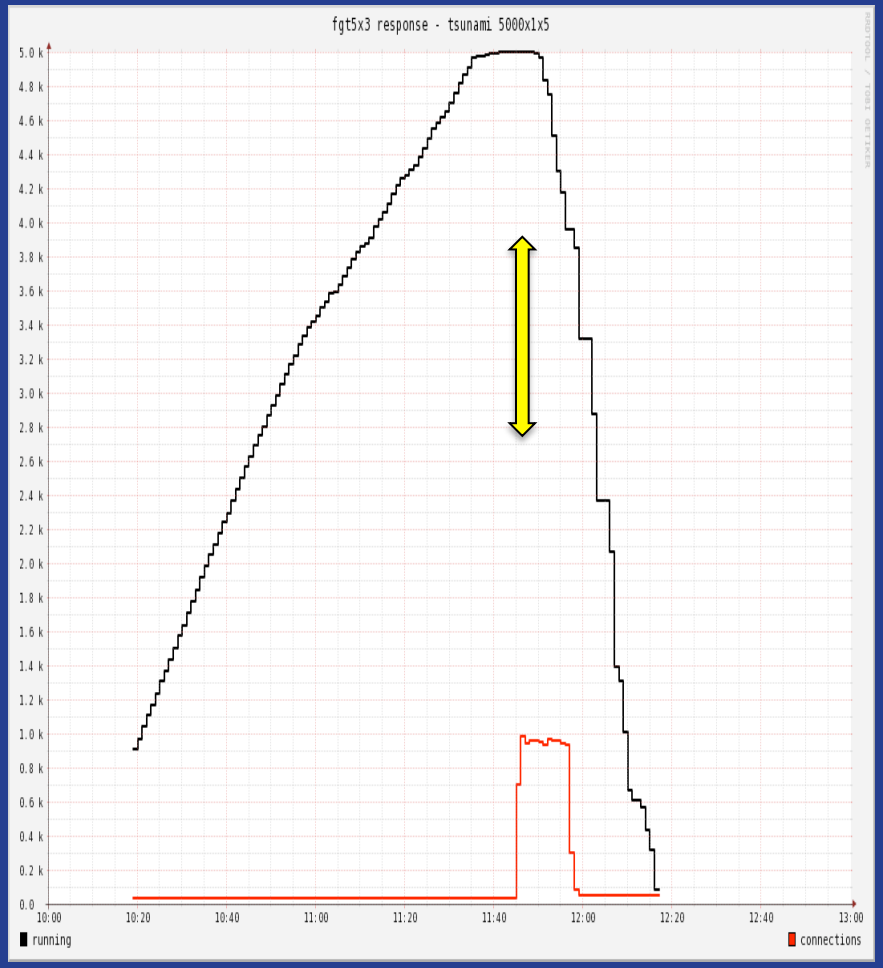
Comparison of Old & New Protocol



Previous SAZ Server Performance

- Black = number of condor jobs,
- Red = number of saz network connections.

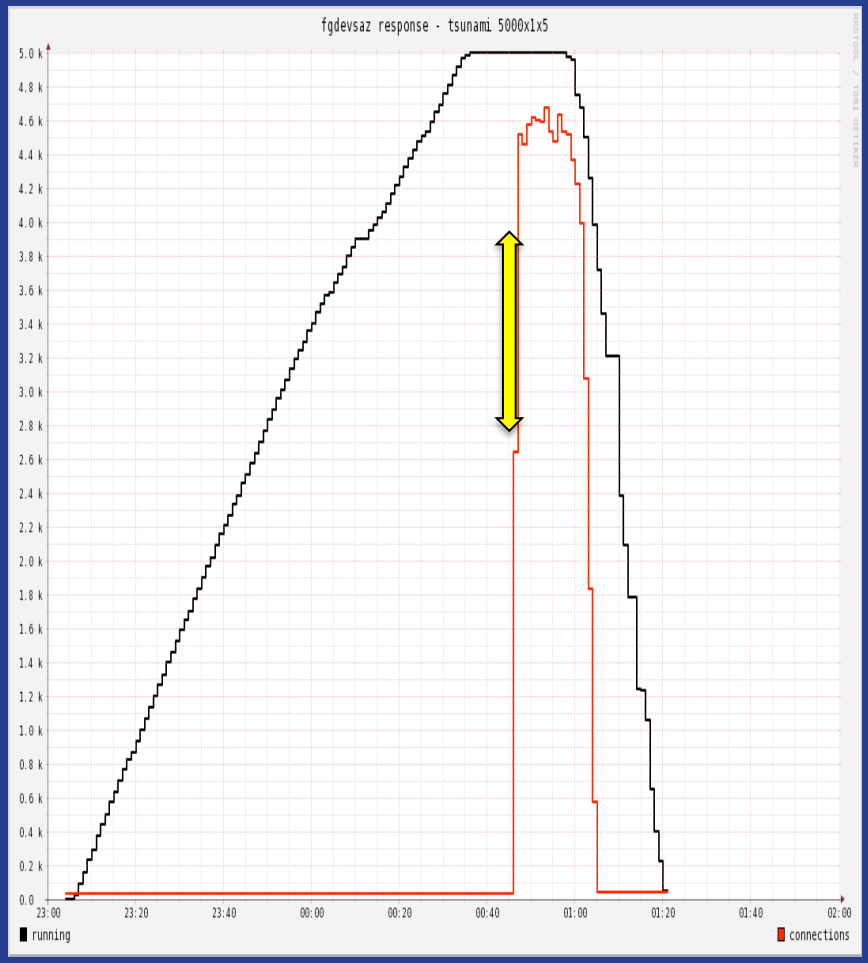
- Trigger @ 11:45:12
- Failures start @ 11:45:19
- 25,000 Authorizations
- 14,183 Success
- 10,817 Failures
- Complete @ 11:58:28
- Elapsed time 13m 16s



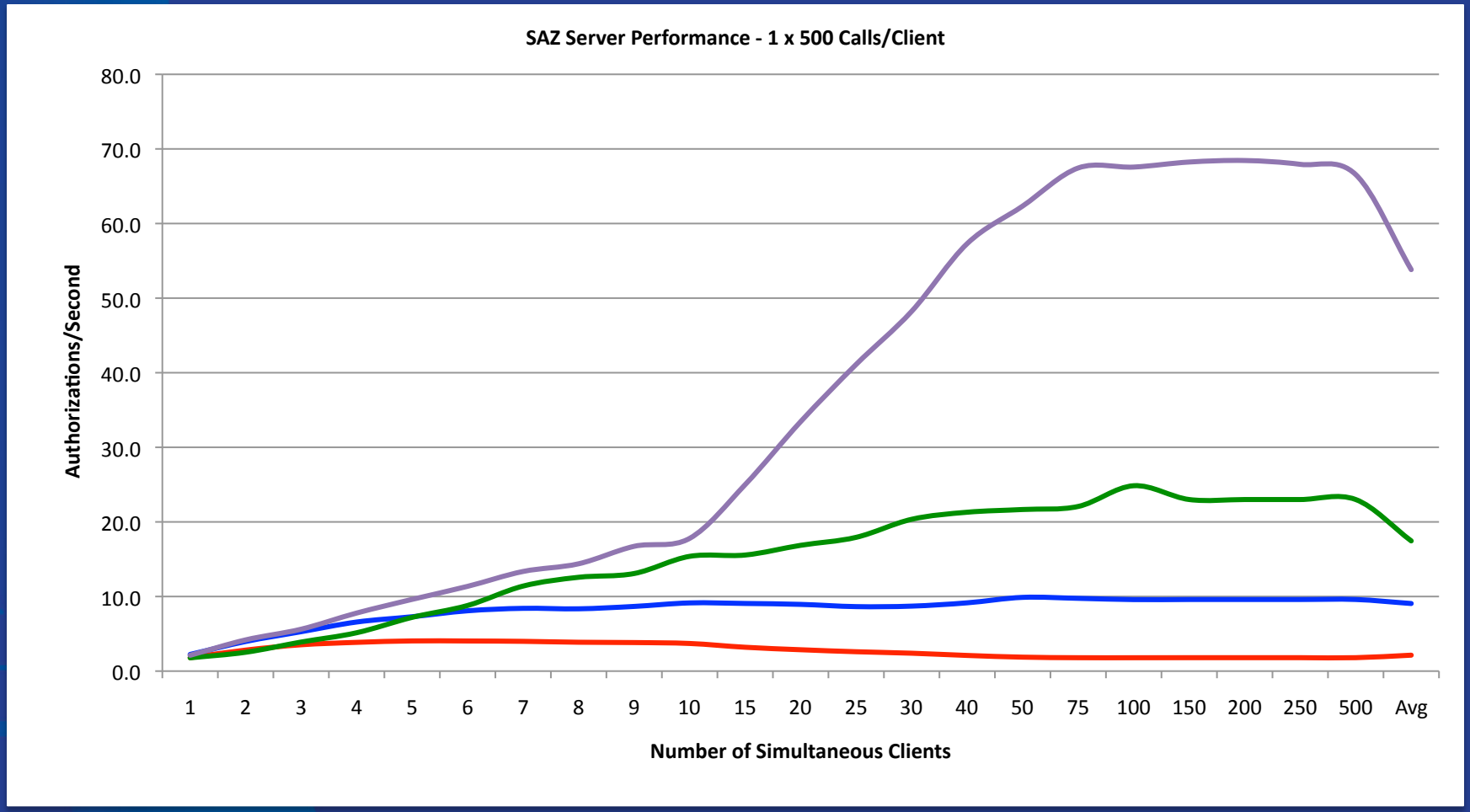
Current SAZ Server Performance

- Black = number of condor jobs,
- Red = number of saz network connections.

- Trigger @ 00:46:20,
- 25,000 Authorizations,
- 25,000 Success,
- 0 Failures,
- Complete @ 01:05:03,
- Elapsed time = 18m 43s,
- 22.26 Authorizations/sec.



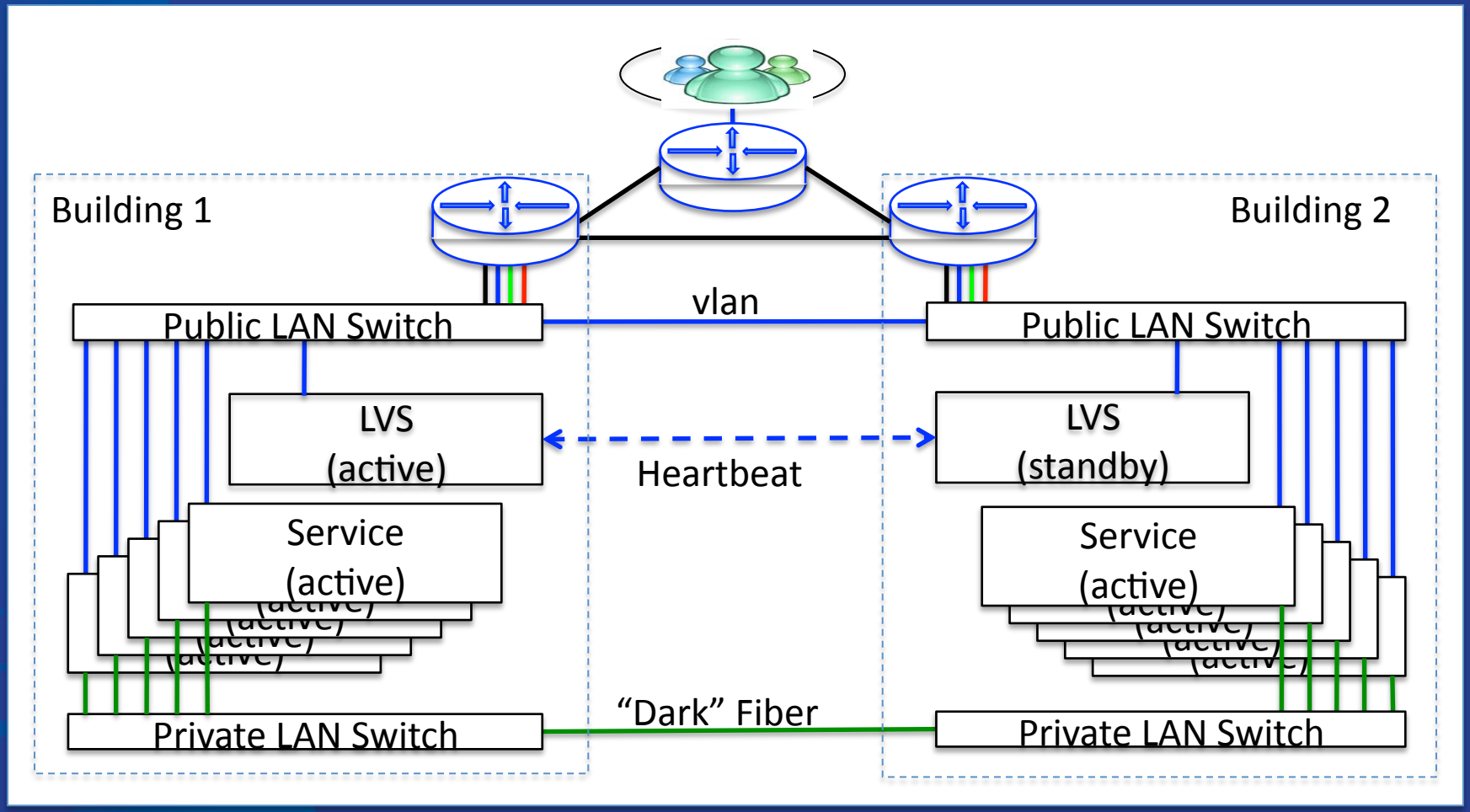
Multiple Client SAZ performance Old (red) vs. Current (purple)



FermiGrid-HA2 Project

- At present, the FermiGrid machines are all in same building on the same power and network infrastructure,
  - Vulnerable to building issues (4X in past ~year),
  - Vulnerable to network issues (6X in past ~quarter).
- The goal of the FermiGrid-HA2 project is to spread the systems and services between two buildings to lessen the chance of network cut or power outage disrupting all service,
- High availability:
  - Web services: LVS active/active with MySQL back end,
  - File systems: DRBD + Heartbeat, (MyProxy and Gratia now, Compute Element soon),
  - Native HA – Condor collector/negotiator.

FermiGrid-HA2 Network



FermiGrid-HA2 Rack Layouts

- Two "almost identically" configured racks,
  - One with 120VAC power distribution,
  - One with 208VAC power distribution,
- Both currently located in FCC1 computer room,
- Waiting on some equipment deliveries, installations and network configuration changes,
- 1st rack will be moved to FCC2 computer room,
- 2nd rack will be moved to GCCB computer room.

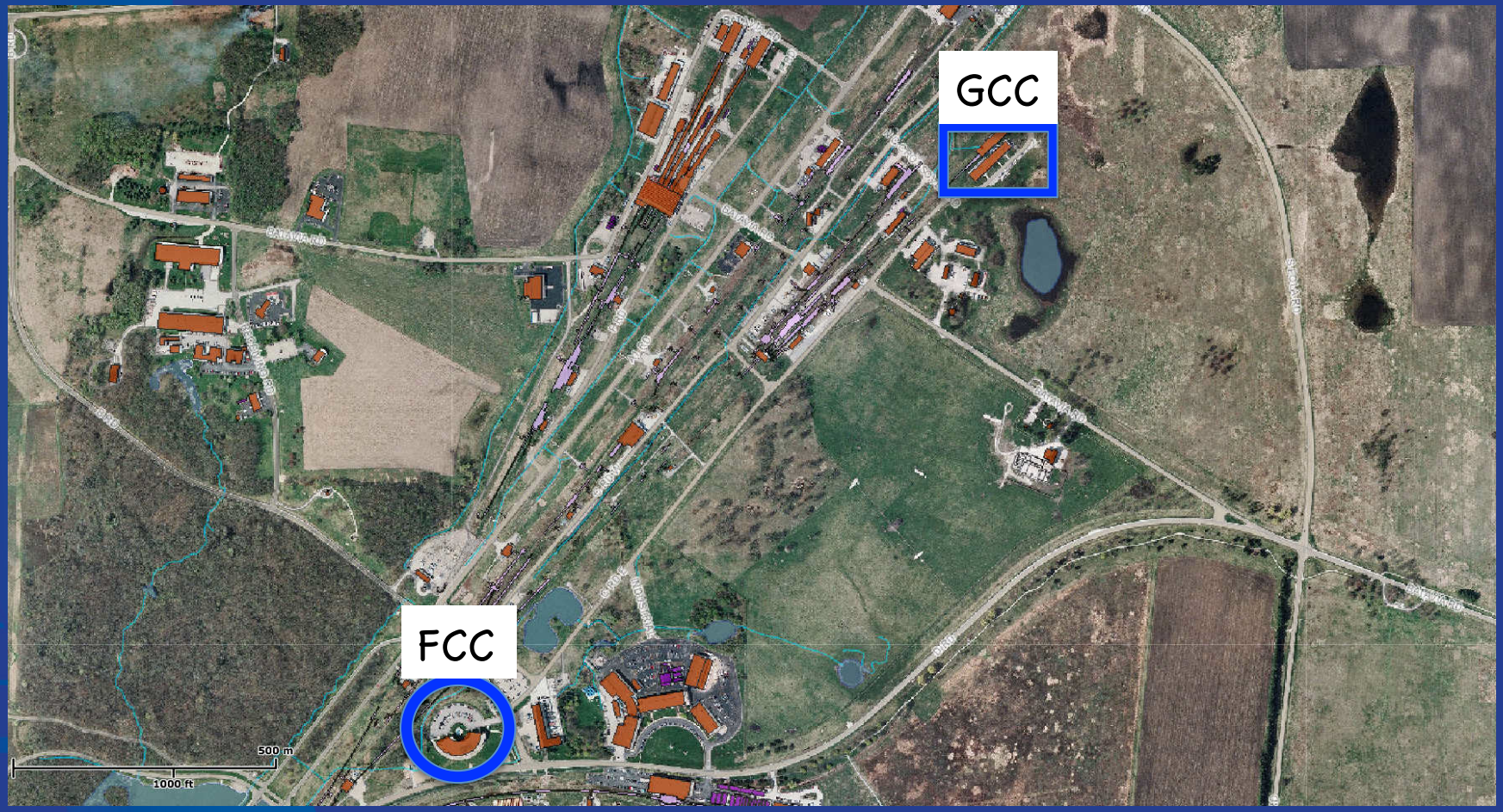
FermiGrid-HA2 Rack Front	Rack "U"	FermiGrid-HA2 Rack Rear
	42	Cisco Nexxus / 2960G Public LAN Switch
blank - 2U	41	Cyclades AlterPath Console Server 16
fnpcsrv8 / blank - 1U	40	fnpcsrv8 / blank - 1U
fnpcsrv5 / fnpcsrv9	39	fnpcsrv5 / fnpcsrv9
fnpcsrv3 / fnpcsrv4	38	fnpcsrv3 / fnpcsrv4
blank - 1U	37	blank - 1U
	36	
d0osgsrv1 / d0osgsrv2	35	d0osgsrv1 / d0osgsrv2
blank - 2U	34	
	33	blank - 2U
fcdfsrv3 / fcdfsrv4	32	
	31	fcdfsrv3 / fcdfsrv4
fcdfsrv1 / fcdfsrv2	30	
	29	fcdfsrv1 / fcdfsrv2
fcdfsrv0 / fcdfsrv5	28	
	27	fcdfsrv0 / fcdfsrv5
blank - 2U	26	Cyclades PM10-L30A 120VAC
	25	Cyclades PM10-L30A 120VAC
Display / Keyboard / Mouse	24	Display / Keyboard / Mouse
Raritan MasterConsole MCCAT116 KVM	23	Omniview PS3 16 port KVM
blank - 1U	22	Linksys SR2024 Private LAN Switch
Slave KDC - Sun Netra X1	21	APC Transfer Switch
blank - 1U	20	blank - 1U
	19	
	18	
gratia12 (FCC) / gratia13(GCC)	17	gratia12 (FCC) / gratia13(GCC)
	16	
	15	
gratia10 (FCC) / gratia11 (GCC)	14	gratia10 (FCC) / gratia11 (GCC)
blank - 1U	13	blank - 1U
	12	
ress01 / ress02	11	ress01 / ress02
	10	
fermigrid5 / fermigrid6	9	fermigrid5 / fermigrid6
	8	
fermigrid2 / fermigrid3	7	fermigrid2 / fermigrid3
	6	
fermigrid1 / fermigrid4	5	fermigrid1 / fermigrid4
	4	
fermigrid0 / fermigrid7	3	fermigrid0 / fermigrid7
	2	Cyclades PM10-L30A 120VAC
blank - 2U	1	Cyclades PM10-L30A 120VAC

Vertical PDU - APC 7941 208VAC

Vertical PDU - APC 7941 208VAC

FermiGrid-HA2 Pictures

Geographical Redundancy



Conclusions

- Good policy planning gave FermiGrid an extensible architecture that we have been able to scale,
- More than 210,000,000 CPU-Hours delivered thus far, most in Open Science Grid,
- Testing at very large scale is key to operating at large scale,
- Expect completion of FermiGrid-HA2 by late May 2011:
  - Will give us resilience against single building failure,
  - FCC power outage currently scheduled for June 2011 will be an important test.
- Ongoing planning and testing is key to reliable service delivery.

Fin

- Any Questions?